CAS Custom Services℠

# How data limitations hold back AI in pharma R&D — and how organizations can reverse the trend

CAS
A division of the
American Chemical Society

# Table of Contents

# The AI promise vs. the data reality

Pharmaceutical organizations have invested aggressively in artificial intelligence (AI) to accelerate insight generation, improve candidate prioritization, and shorten discovery cycles. Despite the momentum, many struggle to translate this spend into measurable impact, with AI initiatives often stalling before they deliver sustainable value.

The reason is not the AI itself, but the data.

This white paper examines the knowledge barriers that limit AI-driven R&D workflows from delivering measurable value, and outlines how organizations can address data readiness at the enterprise level to increase return on AI invesvent.

# Why pilot AI initiatives deliver, but scaling stalls

AI in the pharmaceutical industry often performs well in experimental settings, where the scope is limited and conditions can be manually controlled. However, **nearly two-thirds of organizations** struggle to scale AI initiatives.

The reality is that operating in pilot settings doesn't reflect the complexity of deploying AI models at enterprise scale. As AI in pharma expands beyond pilots, models encounter a fundamentally different scientific data environment, introducing a new level of operational constraints impacting data integration, interoperability, and governance (**Table 1**).

**Table 1**. How AI data requirements change when moving from pilot to enterprise scale.

| AI in pilot environments typically operates with: | AI at enterprise scale requires: |
|---|---|
| Tightly curated and scoped datasets. | Heterogeneous datasets curated from internal and external sources. |
| Minimal data integration. | Cross-functional data integration for enterprise-wide interoperability. |
| One-off data preparation. | Continuous scientific data cleaning, harmonization, and curation. |
| Single therapeutic context. | Multi-domain interoperability (biology, medicinal chemistry, and pharmacology). |
| Heavy reliance on manual scientific oversight. | Built-in, automated data governance. |

At enterprise scale, models must operate within more complex, inconsistent, and variable scientific data environments. Without the manual controls and curation typical of pilot settings, data-related issues compound, putting AI-driven R&D workflows at risk.

# The hidden cost of data fragmentation

The value of AI in pharma R&D workflows depends on high-quality, connected information. However, scientific data scattered across internal and external sources remains largely unstandardized and disconnected. This fragmentation arises at multiple levels, introducing a persistent operational burden that compounds as R&D programs scale (**Table 2**).


**Table 2**. Fragmentation patterns in scientific data and impacts on AI initiatives.

| Fragmentation pattern | Operational impact on AI initiatives |
|---|---|
| **Inconsistent naming conventions and identifiers**<br><br>Examples: SMILES vs. InChI identifiers; target names vs. UniProt accessions. | Delays R&D teams in building coherent datasets, slowing model validation and deployment. |
| **Conflicting results and metadata**<br><br>Examples: Conflicting $EC_{50}$ and $IC_{50}$ values; unit discrepancies across sources. | Requires extensive manual reconciliation, which increases error risk and weakens confidence in model outputs. |
| **Siloed internal systems**<br><br>Examples: LIMS, ELNs, lab journals, and disconnected legacy systems. | Limits models' access to organizational intelligence, which reduces scope and contextual relevance. |
| **Scattered information across external databases**<br><br>Examples: Public and proprietary databases covering biological mechanisms, drug information, and pharmacology data. | Delays model training as teams manually validate and align external inputs. |

Over time, repeated data-preparation cycles increase costs, slow time to insight, and erode confidence in AI-driven outputs. Portfolio decisions are then delayed as teams wait for reliable evidence to advance programs with confidence.

# Three AI-ready data essentials to propel pharma R&D workflows forward

As AI becomes embedded in pharma R&D workflows, data must support repeatable, enterprise-level use without constant manual intervention. This shift requires internal scientific data to meet three essential conditions:

## Scientific accuracy and integrity

AI models fueled by incomplete or biased data produce unreliable outputs that undermine trust in AI-driven workflows. In pharmaceutical R&D, where early decisions shape candidate prioritization and development timelines, even minor gaps or integrity issues can affect downstream approvals and compromise patient safety.

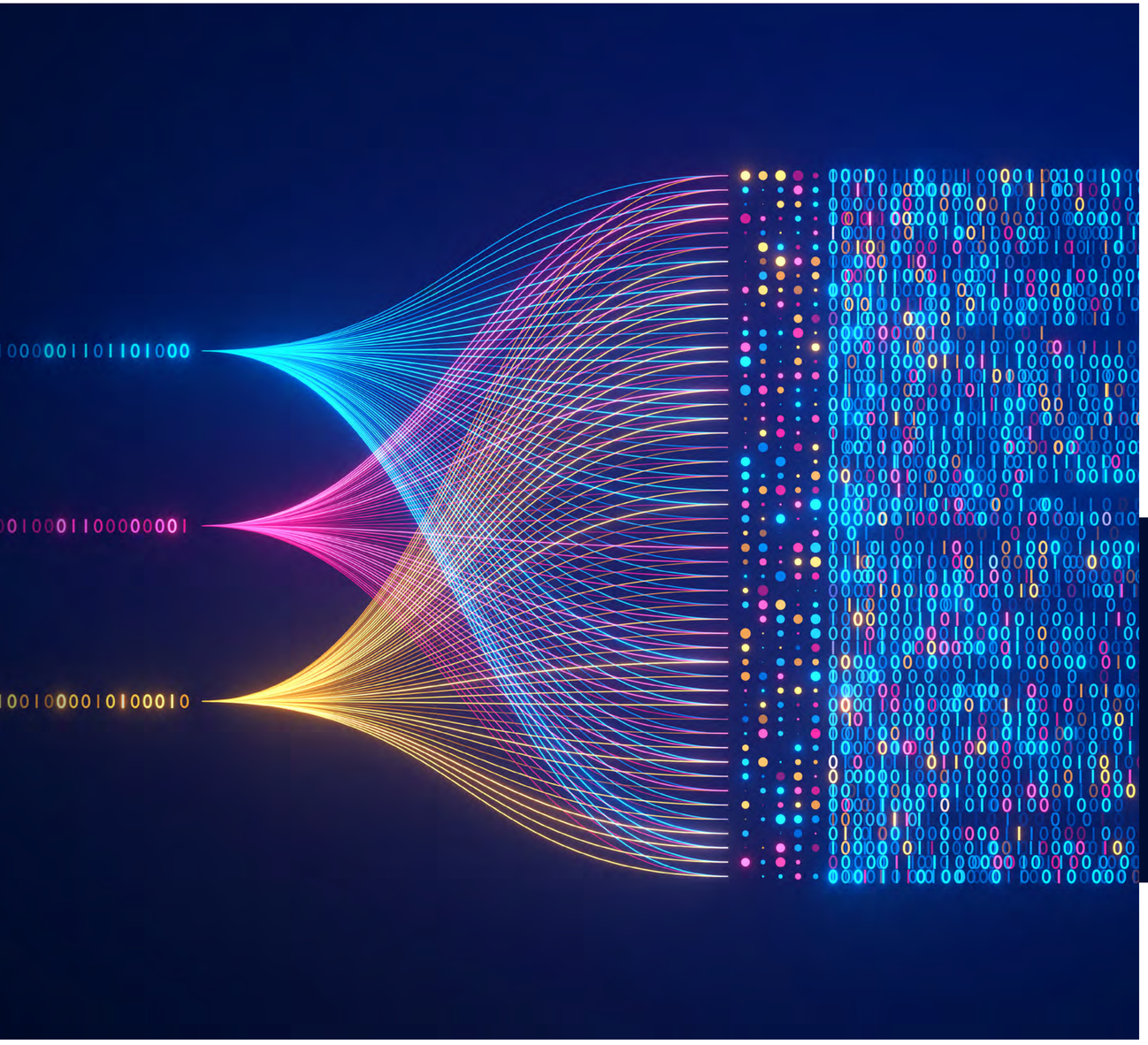
Pharmaceutical organizations must ensure scientific accuracy and integrity to generate reliable AI-driven insights that support an informed, defensible R&D strategy.

To generate trustworthy AI outcomes, scientific data must be:

**Grounded** in validated scientific evidence from reliable sources.

**Contextualized** with experimental conditions, interpretation cues, and relevant data variables.

**Representative** across data types, formats, and sources.



## Consistent data across internal and external sources

When data is inconsistent, teams spend more time reconciling inputs rather than advancing AI initiatives, which slows development cycles and increases R&D costs. More critically, misaligned evidence degrades dataset quality, limiting confidence in AI-driven outputs and introducing uncertainty across candidate progression and portfolio decisions.

As noted previously, teams and departments often operate in functional silos, creating internal inconsistencies that ripple through datasets and undermine the predictive power of AI. These challenges intensify when internal data is combined with external scientific information to add breadth and context. Misalignment across external sources and internal systems obscures the scientific landscape, which impacts the quality of AI-driven insights and limits their strategic value.

To streamline cross-source alignment and strengthen AI outputs, organizations need a robust data foundation defined by three core knowledge management building blocks:

**Standardized data formats, structures, units, and metadata** to support reproducibility and interoperability across R&D functions.

**Harmonized identifiers and nomenclatures for compounds, targets, pathways, and diseases** to streamline data integration and reuse.

**Aligned scientific evidence across internal and external sources** to eliminate conflicting inputs and improve data usability.


## Machine-ready data with built-in governance

Without machine-ready data, AI cannot ingest, analyze, or reuse information without heavy manual preparation. For AI to scale beyond isolated use cases, data must be maintained in structured, machine-readable formats that support automation and repeatable model execution. At scale, this same structure must support interoperability across domains, allowing chemistry, biology, safety, and ADME/Toxicity data to be combined within a unified analytical environment rather than remaining confined to siloed systems.

Data governance, including robust data lineage, versioning, and traceability, is equally important. When governance is flawed, results are hard to audit, and confidence in AI-driven outputs erodes. Built-in governance ensures models are trained on the appropriate data, outputs can be traced back to their source, and AI-driven decisions remain reproducible and auditable over time.

Together, these capabilities create a shared scientific data foundation that enables reliable, reproducible model outputs and supports enterprise interoperability across R&D. With aligned data environments, AI can be deployed consistently across therapeutic areas and program teams rather than rebuilt for each use case.

# Enable science-smart AI with CAS Custom Services

The burden of data readiness often falls on high-value R&D teams, who end up spending time reconciling data, fixing inconsistencies, and filling gaps, rather than advancing programs. CAS Custom Services takes on the complexity of data readiness and maintenance to help pharmaceutical organizations scale AI initiatives while decreasing the burden on R&D teams.

AI outcomes depend on model design, training, and how outputs are ultimately applied. CAS focuses on strengthening the quality, structure, and usability of the scientific data that underpins AI-driven R&D workflows.

## From internal data overload to AI-ready assets

Combining deep scientific expertise with advanced knowledge management capabilities, CAS Custom Services transforms fragmented internal information into strategic, AI-ready assets that support enterprise-scale operations.

Our data management and scientific experts handle the heavy lift by:

- **Digitizing legacy documents** to convert latent information into accessible and actionable digital assets. This includes:
  - Lab journals, research reports, and formulation data.
  - Historical clinical trial and patient records.
  - Regulatory submissions, approval documents, and compliance reports.
  - Patents, trademarks, and proprietary research documents.
- **Harmonizing internal data** to align formats, terminology, and structure across historical and current datasets, enabling reuse at scale.
- **Standardizing data collection systems** with clear entry guidelines and quality check protocols to ensure data reliability from the point of entry.
- **Connect information across silos** through unified knowledge management platforms and custom search capabilities, ensuring AI systems and teams can access and operate on the same trusted information.

By addressing internal data readiness, pharmaceutical organizations can:

Prevent costly rework and delays across development.

Break down data barriers that limit AI performance at enterprise scale.

Extend the value of R&D investment without adding overhead.

## Boost AI performance with trusted, external scientific data

- **CAS Intelligence Hub: Connecting internal and external intelligence.**

  The CAS Intelligence Hub is a cloud-based platform that blends high-quality content from the CAS Content Collection™ with expert curation to strengthen internal data ecosystems. By ingesting, harmonizing, and connecting proprietary information with authoritative external science, CAS enables organizations to integrate cross-domain intelligence and generate confident, AI-ready insights that support enterprise-scale R&D workflows.

**CAS human-curated data has been shown to boost prediction accuracy by more than 30%.**

- **CAS BioFinder®: Zero in on drug discovery data.**

  CAS BioFinder offers an integrated view of chemical, biological, and pharmacological data curated by CAS scientists from leading journals, databases, and patents (**Figure 1**).

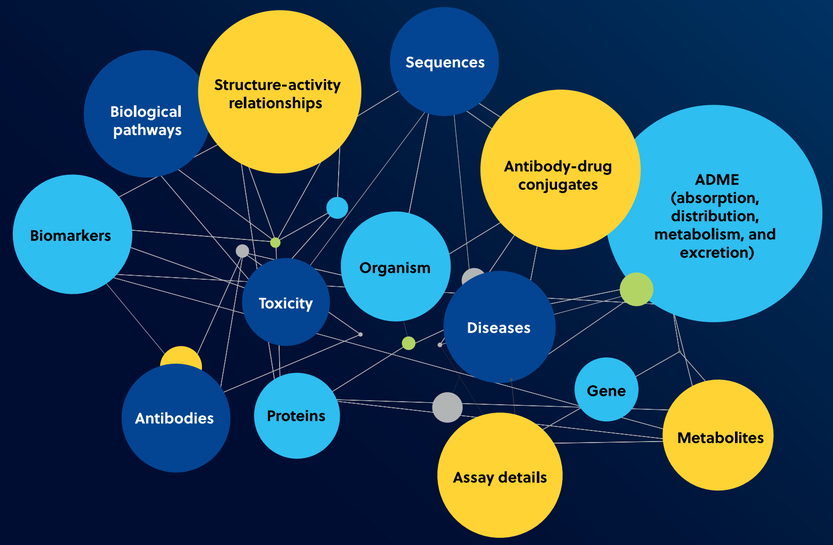



**Figure 1.** Breadth of drug discovery-relevant information available in CAS BioFinder.

By combining trusted content with advanced analytical capabilities, CAS BioFinder enables pharma R&D teams and AI models to focus on high-value data that can be reused directly within in-house models for faster, more confident drug discovery decisions.

Using external inputs to strengthen internal data foundations, pharmaceutical organizations can:

Access a more unified view of relevant scientific landscapes.

Fill data gaps and reduce bias for reliable AI-driven decisions.

Enhance AI model performance without adding complexity.

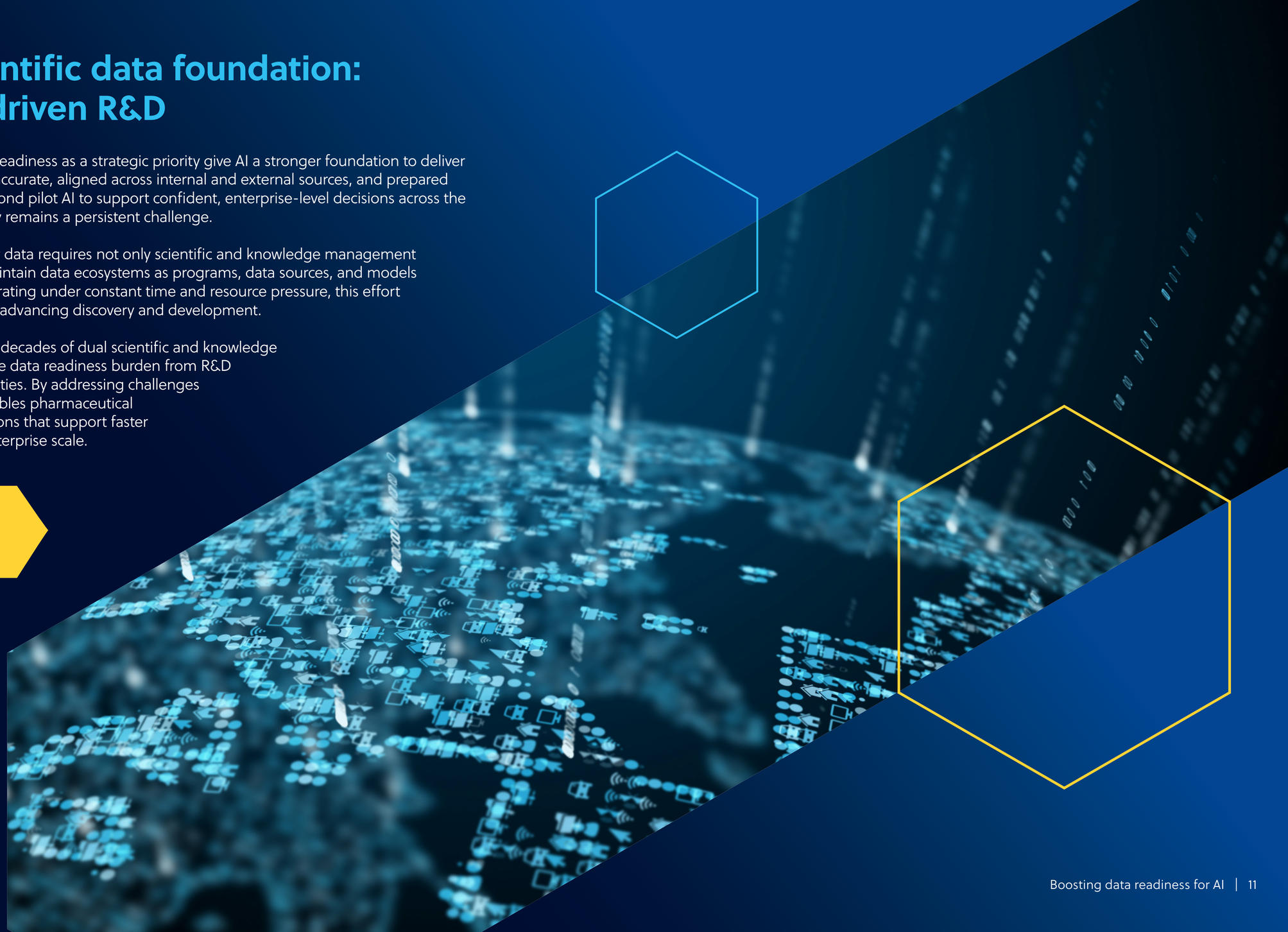
# A high-quality scientific data foundation: The catalyst of AI-driven R&D

Pharmaceutical organizations that treat data readiness as a strategic priority give AI a stronger foundation to deliver sustained R&D value. When scientific data is accurate, aligned across internal and external sources, and prepared for machine use, organizations can move beyond pilot AI to support confident, enterprise-level decisions across the R&D pipeline. However, getting data AI-ready remains a persistent challenge.

Converting relevant information into AI-ready data requires not only scientific and knowledge management expertise, but also ongoing investment to maintain data ecosystems as programs, data sources, and models evolve. For pharmaceutical organizations operating under constant time and resource pressure, this effort strains R&D teams and pulls focus away from advancing discovery and development.

Through dynamic data solutions that harness decades of dual scientific and knowledge management expertise, CAS can decrease the data readiness burden from R&D teams, freeing up capacity for core R&D priorities. By addressing challenges across internal and external sources, CAS enables pharmaceutical organizations to build AI-ready data foundations that support faster deployment and sustained R&D impact at enterprise scale.

**Find out how we can help you get your data AI-ready. Contact us today: [cas.org/contact](cas.org/contact)**



CAS connects the world's scientific knowledge to accelerate breakthroughs that improve lives. We empower global innovators to efficiently navigate today's complex data landscape and make confident decisions in each phase of the innovation journey. As a specialist in scientific knowledge management, our team builds the largest authoritative collection of human-curated scientific data in the world and provides essential information solutions, services, and expertise. Scientists, patent professionals, and business leaders across industries rely on CAS to help them uncover opportunities, mitigate risks, and unlock shared knowledge so they can get from inspiration to innovation faster. CAS is a division of the American Chemical Society.

**Connect with us at cas.org**



SVCWHP102863ENG

CAS
A division of the
American Chemical Society